

Brief Report

Not peer-reviewed version

SARS-CoV-2 and the CGG-CGG Furin Site Genetic Fingerprint: Five Years Later

[Antonio R. Romeu](#) *

Posted Date: 6 March 2025

doi: 10.20944/preprints202503.0413.v1

Keywords: SARS-CoV-2; furin site; arginine codon usage bias; bioinformatics



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Brief Report

SARS-CoV-2 and the CGG-CGG Furin Site Genetic Fingerprint: Five Years Later

Antonio R. Romeu

Biochemistry and Molecular Biology. University Rovira i Virgili. Tarragona. Spain; antonioramon.romeu@iubilo.urv.cat

Abstract: The key evolutionary step leading to the pandemic virus was the acquisition of the furin cleavage motif at the S protein S1/S2 junction. This insertion led to a gain of function for SARS-CoV-2, in which the virus's S protein became a substrate protein for human furin. The corresponding 12 nucleotide fragment inserted into the S gene in a SARS-CoV-2 precursor included the CGG-CGG genetic fingerprint coding the furin arginine pair. The arginine CGG codon was (still is) rare in the virus, even more two CGGs in a row. Afterwards the probable human origin of that motif has been proposed (BMC Genomic Data 24:71, 2023). Synonymous base substitutions or arginine codon usage bias at the CGG-CGG fingerprint was one of the evidences supporting the hypothesis. Based on 2025 SARS-CoV-2 isolates the aim of this work is follow the evolution of the furin site arginine pair code. From GISAID database 17,506 SARS-CoV-2 complete genomes were downloaded, with collection dates from January 1, 2025 to February 18, 2025. Using Perl programs the S gene sequences were retrieved. 62 out of 15,390 (0.4028%) S-protein sequences showed arginine codon usage bias at the S gene CGG-CGG fingerprint. The SARS-CoV-2 lineage distribution of the 2025 sample is shown. The XEC (44.5%) and KP.3.1.1 (13.8%) lineages were the majority. Lineage KP.3.1.1 was also the majority in CGG-CGG codon usage bias analyses, grouped into two main population groups of origin Japan and Canada. In the 2025 working sample 125 out of 1,620 (7.71%) Japan and 47 out of 4,793 (0.98%) Canada Ontario KP.3.1.1. isolates showed CGG-CGG optimization. The results shown are in agreement with previous studies, although in large samples the percentage (probability) of SARS-CoV-2 S gene furin site arginine codon optimization appears weak, it increases significantly when focusing on specific lineages or population groups.

Keywords: SARS-CoV-2; furin site; arginine codon usage bias; bioinformatics

Brief Report

The key evolutionary step leading to the pandemic virus was the acquisition of the furin cleavage motif at the S protein S1/S2 junction. In the first SARS-CoV-2 clinical isolates it was proline (P), arginine (R), arginine and alanine (A) (PRRA) [1]. The corresponding 12 nucleotide fragment inserted into the S gene in a SARS-CoV-2 precursor included the CGG-CGG genetic fingerprint coding the furin arginine pair. The arginine CGG codon was (still is) rare in the virus, even more two CGGs in a row [2–5]. Afterwards the probable human origin of that motif has been proposed [6]. Synonymous base substitutions or arginine codon usage bias at the CGG-CGG fingerprint was one of the evidences supporting the hypothesis and have been reported [6–9]. Based on 2025 SARS-CoV-2 isolates the aim of this work is follow the evolution of the furin site arginine pair code.

Furin Basics

Many proteins are synthesized as non-active precursors, which are subsequently converted to the active form. A mechanism of post-translational modification is through proteases, which are molecular scissors that cut or eliminate part of the non-active precursor to make it an active protein. In human and many organisms,

furin is one of these proteases which acts in the secretory pathway. Technically, furin is a member of the subtilisin-like protein convertase family [10].

Therefore, there are many proteins (or non-active precursors) which are furin substrates, i.e., they have an active site (furin site) that allow “protein-furin” interaction. In any furin substrate protein the furin cleavage site encompasses a 20 amino acid residues fragment (aa1, ..., aa20) that are designated by a specific nomenclature system (P14, P13, ..., P2, P1, P1',P2', ..., P6') (Figure 1). The specific cleavage locus is between positions P1 and P1'. P1 is a strict conserved arginine residue. P5-P1 is the core positively charged (polybasic) motif. P14-P6 and P2'-P6' are small and hydrophilic residues that configure flanking polar side chains providing weak interactions with furin polar surface [11].

	P14	P13	P12	P11	P10	P9	P8	P7	P6	P5	P4	P3	P2	P1	P1'	P2'	P3'	P4'	P5'	P6'	
BANAL-103	A	S	Y	Q	T	Q	T	N	S	-	-	-	-	R	S	V	A	S	Q	S	682
BANAL-236	A	S	Y	Q	T	Q	T	N	S	-	-	-	-	R	S	V	A	S	Q	S	682
RaTG13	A	S	Y	Q	T	Q	T	N	S	-	-	-	-	R	S	V	A	S	Q	S	687
BANAL-52	A	S	Y	Q	T	Q	T	N	S	-	-	-	-	R	S	V	A	S	Q	S	687
SARS-CoV-2 (Wuhan-Hu-1)	A	S	Y	Q	T	Q	T	N	S	P	R	R	A	R	S	V	A	S	Q	S	691
SARS-CoV-2 (WH04)	A	S	Y	Q	T	Q	T	N	S	P	R	R	A	R	S	V	A	S	Q	S	691
	*	*	*	*	*	*	*	*	*					*	*	*	*	*	*	*	

Fragment of protein multiple sequence alignment covering the SARS-CoV-2 S protein polybasic furin cleavage site. Line at the top indicates the positions (P41-P6' nomenclature) of the canonical structure of a furin site [11]. The cleavage locus is between positions P1 and P1'. The core regions is between positions P5-P1 and there are two flanking solvent accessible regions: P7-P14 and P2'-P6'. The fragment of the protein multiple sequence alignment includes Laos bat Rhinolophus coronaviruses BANAL-52 (GISAID, EPI_ISL_4302644: 21512-25321), BANAL-103 (GISAID, EPI_ISL_4302645: 21498-25294), BANAL-236 (GISAID, EPI_ISL_4302647: 21538-25344), Bat coronavirus RaTG13 (GenBank: QHR63300.2 coded by MN996532.2: 21560..25369 genome) and the reference SARS-CoV-2 sequences (isolates Wuhan-Hu-1 and WH04) [16]. The SARS-CoV-2 polybasic insert (PRRA) is denoted in bold. Strictly conserved amino acids are denoted by *. The amino acid position is indicated at the numbers on the right.

Figure 1. SARS-CoV-2 S protein furin site.

Furin and SARS-CoV-2

The role of the human furin in the SARS-CoV-2 biology is during the S protein biosynthesis and maturation [12]. When the newly synthesized viral S protein transits through the Golgi apparatus of the infected cell and because the S-protein has the complete furin site, the human furin cuts into the S1 and S2 subunits, which remain associated. The S protein on the virus therefore consists of two non-covalently associated subunits with different functions: in the new target cell, the S1 subunit binds the ACE2 receptor and the S2 subunit anchors the S protein to the virion membrane and mediates membrane fusion [12].

With the acquisition of the furin polybasic motif the SARS-CoV-2 S-protein became a protein-substrate [13] of the human furin.

Furin Arginine Pair Codon Usage Bias in 2025 SARS-CoV-2 Isolates

Based on the GISAID database, 17,506 SARS-CoV-2 genomes were downloaded, with collection dates from January 1, 2025 to February 18, 2025. From each genome, the region covering the S gene was extracted for analysis. 15,390 out of 17.506 S gene sequences could be used in this work. The encoded S protein sequences were complete and had no ambiguous characters at the core of the furin site (P5-P1). As a result, 62 out of 15,390 (0.4028%) S-protein sequences showed arginine codon usage bias at the S gene CGG-CGG fingerprint (Table 1).

When the SARS-CoV-2 lineages were taken into account the KP.3.1.1 lineage was the majority 42 out of 62 (67,74%) (Table 2), which was grouped into two main geographic regions: Japan and Canada. In this 2025 working sample 125 out of 1,620 (7,71%) Japan and 47 out of 4,793 (0,98%) Canada Ontario KP.3.1.1. isolates showed CGG-CGG codon usage optimization.

Table 3 shows the SARS-CoV-2 lineage distribution of viruses whose complete genomes were downloaded from the GISAID database in the initial working sample. XEC (7,792 out of 17,506 viruses; 44.5%) and KP.3.1.1 (2,414 out of 17,506; 13.8%) SARS-CoV-2 lineages were the majority.

Table 1. Synonymous base substitution (codon usage bias) at the SARS-CoV-2 arginine pair (CGG–CGG) S gene furin cleavage site.

• sorted by collection date

GISAID_epi_isl id	Region	Country	Division	Length*	Lineage	Collect. Date	RR**	RR_Coding***	Furin_Cleavage_Site**
NI_045512.2 (GenBank)	Asia	China	Hubei (Wuhan)	29903	B	2019-12-26	RR-683	CGGCGG-23659	672- ASYQTQTNSPRRRARSVASQS- 691
EPI_ISL_406801	Asia	China	Hubei (Wuhan)	29872	A	2020-01-30	RR-682	CGGCGG-23594	671- ASYQTQTNSPRRRARSVASQS- 690
EPI_ISL_19679698	North America	Canada	British Columbia	29553	KP.3.1.1	2025-01-01	RR-672	CGGCGT-23531	661- ASYQTQTKSRRRRARSVASQS- 680
EPI_ISL_19700434	Asia	Israel		29401	LF.7.1.3	2025-01-02	RR-673	CGGCGA-23252	662- ASYQTQTKSRRRRARSVASQS- 681
EPI_ISL_19684930	Asia	Japan	Hyogo	29749	KP.3.1.1	2025-01-02	RR-672	CGGCGT-23531	661- ASYQTQTKSRRRRARSVASQS- 680
EPI_ISL_19682951	North America	USA	Arizona	29721	XEC	2025-01-02	RR-673	CGGCGT-23538	662- ASYQTQTKSRRRRARSVASQS- 681
EPI_ISL_19666094	North America	USA	Illinois	29690	XEC	2025-01-02	RR-673	AGGCGG-23516	662- ASYQTQTKSRRRRARSVASQS- 681
EPI_ISL_19663311	North America	Canada	Ontario	29722	KP.3.1.1	2025-01-04	RR-674	CGGCGA-23513	663- XXXXXQTKSRRRRARSVASQS- 682
EPI_ISL_19682995	North America	USA	New York	29698	XEC	2025-01-04	RR-673	CGTCGG-23538	662- ASYQTQTKSRRRRARSVASQS- 681
EPI_ISL_19681312	Europe	Germany	North Rhine-Westphalia	29706	XEC	2025-01-06	RR-673	CGGCGT-23510	662- XXXXXQTKSRRRRARSVASQS- 681

EPI_ISL_19666768	Asia	Singapore		29569	LF.7.3.1	2025-01-06	RR-673	CGACGG-23310	662-ASYQTXTKSRRRARSVASQS-681
EPI_ISL_19706176	Asia	South Korea		29795	KP.3.1.1	2025-01-06	RR-668	CGGCGT-23569	657-ASYQTQTKSRRRARSVASQS-676
EPI_ISL_19675573	North America	Canada	Ontario	29716	KP.3.1.1	2025-01-07	RR-672	CGGCGT-23507	661-XXXXXQTKSRRRARSVASQS-680
EPI_ISL_19683050	North America	USA	New York	29698	XEC	2025-01-07	RR-673	CGTCGG-23538	662-ASYQTQTKSRRRARSVASQS-681
EPI_ISL_19676763	Oceania	Australia	Queensland	29721	XEC	2025-01-08	RR-673	CGTCGG-23538	662-ASYQTQTKSRRRARSVASQS-681
EPI_ISL_19684954	Asia	Japan	Hokkaido	29723	KP.3.1.1	2025-01-08	RR-672	CGGCGT-23531	661-ASYQTQTKSRRRARSVASQS-680
EPI_ISL_19701204	Europe	United Kingdom	England	29738	JN.1.8	2025-01-08	RR-672	CGGCGT-23521	661-XSYQTQTKSRRRARSVASQS-680
EPI_ISL_19684105	North America	USA	Colorado	29703	KP.3.1	2025-01-08	RR-672	CGTCGG-23507	661-ASYQTQTKSRRRARSVASQS-680
EPI_ISL_19691313	North America	Canada	Ontario	29716	KP.3.1.1	2025-01-10	RR-672	CGGCGT-23507	661-XXXXXQTKSRRRARSVASQS-680
EPI_ISL_19696722	Asia	Japan	Miyagi	29723	LP.8.1	2025-01-10	RR-672	CGGCGA-23531	661-ASYQTQTKSRRRARSVASQS-680
EPI_ISL_19707783	North America	USA	Hawaii	29703	KP.3.1.1	2025-01-10	RR-672	CGGCGT-23507	661-ASYQTQTKSRRRARSVASQS-680

EPI_ISL_19729037	Asia	Japan	Toyama	29723	KP.3.1.1	2025-01-11	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19693631	North America	Canada	Quebec	29716	KP.3.1.1	2025-01-12	RR-672	CGGCGT-23507	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19684922	Asia	Japan	Hyogo	29723	KP.3.1.1	2025-01-12	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19684921	Asia	Japan	Hyogo	29723	KP.3.1.1	2025-01-12	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19691627	North America	Canada	Ontario	29716	KP.3.1.1	2025-01-14	RR-672	CGGCGT-23507	XXXXXQTKSRRRARSVASQS-661-680
EPI_ISL_19731362	South America	Ecuador	Cotopaxi	29830	JN.1.11	2025-01-14	RR-669	AGGCGG-23572	ASYQTQTKSRRRARSVASQS-658-677
EPI_ISL_19700994	Asia	Japan	Gifu	29723	KP.3.1.1	2025-01-14	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19707721	Asia	Japan	Saitama	29723	KP.3.1.1	2025-01-14	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19697568	Asia	Japan	Tokushima	29723	KP.3.1.1	2025-01-14	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19712229	Oceania	Australia	Queensland	29721	KP.3.3	2025-01-15	RR-673	CGGCGA-23538	ASYQTQTKSRRRARSVASQS-662-681
EPI_ISL_19696525	Oceania	Australia	South Australia	29673	KP.3.3	2025-01-15	RR-672	AGGCGG-23512	ASYQTQTKSRRRARSVASQS-661-680

EPI_ISL_19695415	Europe	Germany	Baden-Wurttemberg	29703	KP.3.1.1	2025-01-15	RR-672	AGGCGG-23507	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19681098	Asia	Japan	Kanagawa	29723	KP.3.1.1	2025-01-15	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19691828	North America	Canada	Ontario	29716	KP.3.1.1	2025-01-16	RR-672	CGGCGT-23507	XXXXXQTKSRRRARSVASQS-661-680
EPI_ISL_19691829	North America	Canada	Ontario	29716	KP.3.1.1	2025-01-16	RR-672	CGGCGT-23507	XXXXXQTKSRRRARSVASQS-661-680
EPI_ISL_19700973	Asia	Japan	Gifu	29723	KP.3.1.1	2025-01-16	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19692038	North America	Canada	Ontario	29716	KP.3.1.1	2025-01-17	RR-672	CGGCGT-23507	XXXXXQTKSRRRARSVVSQS-661-680
EPI_ISL_19692035	North America	Canada	Ontario	29716	KP.3.1.1	2025-01-17	RR-672	CGGCGT-23507	XXXXXQTKSRRRARSVVSQS-661-680
EPI_ISL_19692037	North America	Canada	Ontario	29716	KP.3.1.1	2025-01-17	RR-672	CGGCGT-23507	XXXXXQTKSRRRARSVVSQS-661-680
EPI_ISL_19702183	North America	Canada	Ontario	29722	KP.3.1.1	2025-01-18	RR-674	CGGCGT-23513	XXXXXQTKSRRRARSVVSQS-663-682
EPI_ISL_19693229	Asia	Hong Kong		29706	XDV.1	2025-01-18	RR-673	AGGCGG-23510	ASYQTQTKSRRRARSVASQS-662-681
EPI_ISL_19714873	Oceania	Australia	South Australia	29673	KP.3.3	2025-01-19	RR-672	AGGCGG-23512	ASYQTQTKSRRRARSVASQS-661-680

EPI_ISL_19714872	Oceania	Australia	South Australia	29673	KP.3.3	2025-01-19	RR-672	AGGCGG-23512	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19704613	Asia	Japan	Hyogo	29723	KP.3.1.1	2025-01-19	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19700970	Asia	Japan	Gifu	29723	KP.3.1.1	2025-01-20	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19696435	Asia	Japan	Hokkaido	29723	KP.3.1.1	2025-01-20	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19697628	Asia	Japan	Hokkaido	29723	KP.3.1.1	2025-01-20	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19729193	Asia	Japan	Miyagi	29723	KP.3.1.1	2025-01-20	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19729288	Asia	Japan	Saitama	29723	KP.3.1.1	2025-01-20	RR-672	CGTCGG-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19712131	Asia	Japan	Ehime	29723	KP.3.1.1	2025-01-22	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19702323	North America	Canada	Ontario	29716	JN.1.8	2025-01-23	RR-674	CGGAGG-23507	ASYQTQTKSRRRARSVASQS-663-682
EPI_ISL_19729202	Asia	Japan	Miyagi	29723	KP.3.1.1	2025-01-23	RR-672	CGGCGT-23531	ASYQTQTKSRRRARSVASQS-661-680
EPI_ISL_19716895	North America	USA	New Jersey	29681	JN.1.8	2025-01-24	RR-672	CGTCGG-23507	ASYQTQTKSRRRARSVASQS-661-680

EPI_ISL_19724894	Asia	Japan	Tokyo	29749	KP.3.1.1	2025-01-26	RR-672	CGGCGT-23531	ASYQTQTKSRRRRARSVASQS-661-680
EPI_ISL_19731882	Europe	United Kingdom	England	29749	XEC	2025-01-30	RR-672	CGTCGG-23530	ASYQTQTKSRRRRARSVVSQS-661-680
EPI_ISL_19718119	North America	Canada	Ontario	29703	KP.3.1.1	2025-02-03	RR-672	CGGCGT-23507	ASYQTQTKSRRRRARSVVSQS-661-680
EPI_ISL_19718127	North America	Canada	Ontario	29703	KP.3.1.1	2025-02-03	RR-672	CGGCGT-23507	ASYQTQTKSRRRRARSVVSQS-661-680
EPI_ISL_19718123	North America	Canada	Ontario	29716	KP.3.1.1	2025-02-03	RR-672	CGGCGT-23507	XXXXXQTKSRRRRARSVVSQS-661-680
EPI_ISL_19722193	Asia	Japan	Hokkaido	29723	KP.3.1.1	2025-02-03	RR-672	CGGCGT-23531	ASYQTQTKSRRRRARSVASQS-661-680
EPI_ISL_19731403	Asia	Japan	Saitama	29725	KP.3.1.1	2025-02-03	RR-672	CGGCGT-23533	ASYQTQTKSRRRRARSVASQS-661-680
EPI_ISL_19718129	North America	Canada	Ontario	29703	KP.3.1.1	2025-02-04	RR-672	CGGCGT-23507	ASYQTQTKSRRRRARSVVSQS-661-680
EPI_ISL_19722192	Asia	Japan	Hokkaido	29723	KP.3.1.1	2025-02-05	RR-672	CGGCGT-23531	ASYQTQTKSRRRRARSVASQS-661-680
EPI_ISL_19722186	Asia	Japan	Hokkaido	29723	KP.3.1.1	2025-02-06	RR-672	CGGCGT-23531	ASYQTQTKSRRRRARSVASQS-661-680

Reference SARS-CoV-2 spike glycoprotein sequences [16]: the first two rows highlighted in gray. *: SARS-CoV-2 genome length. **: positions corresponding to the analyzed S protein sequence (see methods). ***: position correspondig to the downloaded SARS-CoV-2 complte genome..

Table 2. SARS-CoV-2 lineage distribution at the 2025 isolates with synonymous base substitution (codon usage bias) at the SARS-CoV-2 arginine pair S gene furin cleavage site.

- Total records 62
- Number of virus lineages 10

Lineage	Percent	Number of Isolates
KP.3.1.1	1.74	42
XEC	0.09	7
KP.3.3	2.07	4
JN.1.8	0.64	3
LF.7.1.3	6.67	1
LF.7.3.1	6.25	1
XDV.1	1.59	1
KP.3.1	0.51	1
JN.1.11	0.35	1
LP.8.1	0.12	1

Table 3. SARS-CoV-2 lineage distribution of viruses whose complete genomes were downloaded from the GISAID database in the initial 2025 working sample.

- Isolates collection date from January 1, 2025 to February 18, 2025
- Total records 17,506
- Number of SARS-CoV-2 lineages 183

Lineage	Percent	Number of Isolates
XEC	44.5105	7792
KP.3.1.1	13.7896	2414
MC.1	5.2839	925
LP.8.1	4.8783	854
LP.8.1.1	4.4899	786
JN.1.8	2.6962	472
JN.1.16.1	2.4106	422
JN.1.16	1.7365	304
JN.1.11	1.6223	284
KP.3.3.2	1.1882	208
KP.3.1	1.1310	198
KP.3.3	1.1025	193
MC.10.1	0.9540	167
JN.1.16.3	0.9482	166
LF.7.2.1	0.7712	135
MB.1.1	0.6626	116
KP.3	0.5655	99
LF.7.1	0.5370	94
LP.8.1.2	0.5027	88
KP.2	0.4913	86
LF.7	0.4513	79
XEC.1	0.4456	78
MC.10.2.1	0.4341	76
MC.1.2	0.3713	65
XDV.1	0.3599	63
XDY	0.3427	60
JN.1	0.2856	50
KP.1.1.3	0.2799	49
LF.7.3	0.2742	48

MC.21.1	0.2399	42
JN.1.40	0.2171	38
JN.1.8.1	0.2114	37
KS.1.1	0.2114	37
LF.7.1.2	0.1942	34
MC.13	0.1942	34
KP.3.3.1	0.1828	32
MC.2	0.1828	32
MC.10.1.6	0.1599	28
MC.1.3	0.1542	27
MC.10.1.1	0.1485	26
JN.11	0.1371	24
MC.13.2	0.1257	22
MC.1.5	0.1200	21
MC.1.6	0.1028	18
MC.24	0.1028	18
KP.1.1	0.0971	17
MC.8.1	0.0971	17
LF.7.3.1	0.0914	16
LP.7	0.0914	16
KP.1	0.0857	15
LF.7.1.3	0.0857	15
LP.5	0.0857	15
MC.9	0.0857	15
KP.1.1.5	0.0800	14
KP.3.2	0.0800	14
MC.10.1.2	0.0800	14
MC.10.2	0.0743	13
JN.1.3	0.0685	12
KP.3.1.4	0.0685	12
MC.1.1	0.0685	12
JN.1.18	0.0628	11
KP.1.1.1	0.0628	11
LF.7.2	0.0628	11
MC.11	0.0628	11
JN.1.13	0.0571	10
BA.2.86.1	0.0514	9
JN.1.37	0.0514	9
LF.7.6.1	0.0514	9
MC.1.4	0.0514	9
MC.33.1	0.0514	9
JN.1.11.1	0.0457	8
JN.1.15	0.0457	8
KP.2.2	0.0457	8
KP.2.3.12	0.0457	8
MC.10	0.0457	8
MC.35	0.0457	8
JN.1.18.6	0.0400	7
KP.2.9	0.0400	7
KP.3.3.3	0.0400	7
MC.1.7	0.0400	7
MC.10.1.4	0.0400	7
MC.28	0.0400	7

PA.1	0.0400	7
XDQ	0.0400	7
JN.1.32	0.0343	6
KP.3.1.3	0.0343	6
KS.1	0.0343	6
LU.2	0.0343	6
MC.16	0.0343	6
MC.19	0.0343	6
MC.8	0.0343	6
Unassigned	0.0343	6
BA.2.86	0.0286	5
JN.1.10	0.0286	5
KP.2.3	0.0286	5
LB.1.1	0.0286	5
LP.9	0.0286	5
MB.1	0.0286	5
MC.10.1.5	0.0286	5
MC.13.1	0.0286	5
MC.31	0.0286	5
MC.4	0.0286	5
BA.3	0.0228	4
KP.5	0.0228	4
LW.1	0.0228	4
MC.13.3	0.0228	4
MC.13.4	0.0228	4
MC.23	0.0228	4
MC.26	0.0228	4
MC.34	0.0228	4
JN.1.20	0.0171	3
JN.1.4	0.0171	3
JN.1.59	0.0171	3
JN.1.9	0.0171	3
JN.10	0.0171	3
KP.2.3.4	0.0171	3
KP.3.2.3	0.0171	3
LP.8.2	0.0171	3
MC.1.3.1	0.0171	3
MC.10.1.3	0.0171	3
MC.13.2.1	0.0171	3
MC.3	0.0171	3
MC.30	0.0171	3
MC.6	0.0171	3
ND.1.1.2	0.0171	3
JN.1.13.1	0.0114	2
JN.1.4.4	0.0114	2
JN.1.9.2	0.0114	2
KP.2.3.6	0.0114	2
KP.3.5	0.0114	2
MC.17.1	0.0114	2
MC.19.1	0.0114	2
MC.2.1	0.0114	2
MC.27	0.0114	2
ND.1.1	0.0114	2

NT.1	0.0114	2
XDQ.1	0.0114	2
BA.1.1	0.0057	1
BA.2	0.0057	1
BQ.1.1	0.0057	1
BQ.1.1.18	0.0057	1
BQ.1.18	0.0057	1
FL.4.2	0.0057	1
HF.1.1	0.0057	1
HV.1	0.0057	1
JG.3	0.0057	1
JN.1.18.3	0.0057	1
JN.1.22	0.0057	1
JN.1.4.6	0.0057	1
JN.1.49	0.0057	1
JN.1.49.1	0.0057	1
JN.1.51	0.0057	1
JN.1.52	0.0057	1
JN.1.53	0.0057	1
JN.1.7.3	0.0057	1
KP.2.14	0.0057	1
KP.2.3.10	0.0057	1
KP.3.1.2	0.0057	1
KP.3.2.5	0.0057	1
KP.4	0.0057	1
KP.4.2	0.0057	1
KS.1.2	0.0057	1
LB.1.3.1	0.0057	1
LF.3.1	0.0057	1
LF.7.4	0.0057	1
LF.7.6	0.0057	1
LP.10.1.1	0.0057	1
LV.1	0.0057	1
MA.1	0.0057	1
MC.14	0.0057	1
MC.17	0.0057	1
MC.21	0.0057	1
MC.24.1	0.0057	1
MC.26.1	0.0057	1
MC.28.1	0.0057	1
MC.28.1.1	0.0057	1
MC.32.1	0.0057	1
MC.33.2	0.0057	1
MC.37	0.0057	1
ML.1	0.0057	1
ML.2	0.0057	1
MM.1	0.0057	1
XDK.3	0.0057	1

The results shown here are in agreement with previous studies [6] that taking all together synonymous base substitutions or arginine pair codon usage bias at SARS-CoV-2 S protein furin site is strongly supported. Although in large samples the percentage (probability) of SARS-CoV-2 S gene

furin site arginine codon optimization appears weak, it increases significantly when focusing on specific lineages or population groups.

Methods

The source of information was the Global Initiative on Sharing Avian Influenza Data (GISAID) database [14,15]. The reference SARS-CoV-2 spike glycoprotein sequences were retrieved from the SARS-CoV-2 reference genomes: (i) isolate Wuhan-Hu-1, GenBank: QHD43416.1 coded by MN908947.3:21563-25384; and (ii) isolate WH04, GenBank: QHR63260.2 coded by MN996528.1:21563-25384 and GISAID, EPI_ISL_406801, genome hCoV-19/Wuhan/WH04/2020: 21551-25370 [16]. A pipeline of scripts in Perl for data management has been created. The rationale of this work was based on the following tasks:

Task 1. Getting sequences. Complete SARS-CoV-2 genomes were downloaded from GISAID database. Obtaining the S gene coding region and S protein sequences required data parse by executing several chained programs. Briefly:

- To retrieve the genome region covering the S gene sequence (positions 20000-26000).
- Using NCBI BLASTn [17,18], identification of the s gene start and end coordinates. Query: flanking regions of reference NCBI S gene, sequence; subject: genomic regions covering the S gene [16].
- Based on these coordinates, to retrieve the S gene region from the downloaded GISAID genome.
- To translate forward three frames of the retrieved S gene region (coding region).
- To identify the correct translation reading frame (no ambiguous characters, no stop signals).
- Based on the proper reading frame, to adjust the S gene coding sequence.

Task 2. Synonymous base substitution at the SARS-CoV-2 furin site arginine pair.

- To identify the furin site arginine pair. For each S protein sequence a 2-position RR-window was run. Using the RRAR pattern it was confirmed that the identified arginine pair corresponded to the arginine pair of the furin site.
- To identify the codon usage of the furin site argini pair. Knowing the positions of the arginine pair in the protein, multiplying by three the respective codons were extracted from the S gene sequences. The cases in which the arginine pair code was not CGG-CGG were recorded.

EMBL Clustal Omega tool [19,20] program was used for multiple sequence alignment.

Acknowledgement: The author acknowledges GISAID database contributors.

Competing interests: The author declares he has no competing interests.

References

1. Kristian G Andersen, Andrew Rambaut, W Ian Lipkin, Edward C Holmes, Robert F Garry. The proximal origin of SARS-CoV-2. Nat. Med. 26:450-452, 2020. PMID: 32284615. doi:10.1038/s41591-020-0820-9.
2. Murat Seyran, Damiano Pizzol, Parise Adadi, Tarek M A El-Aziz, Sk Sarif Hassan, Antonio Soares, Ramesh Kandimalla, Kenneth Lundstrom, Murtaza Tambuwala, Alaa A A Aljabali, Amos Lal, Gajendra K Azad, Pabitra P Choudhury, Vladimir N Uversky, Samendra P Sherchan, Bruce D Uhal, Nima Rezaei, Adam M Brufsky. Questions concerning the proximal origin of SARS-CoV-2. J. Med. Virol. 93(3):1204-1206, 2021. PMID:32880995. doi:10.1002/jmv.26478.
3. Steven Quay, Richard Muller. The science suggests a Wuhan lab leak. The Wall Street Journal. Monday, June 7, 2021.
4. Antonio R. Romeu, Enric Ollé. SARS-CoV-2 and the Secret of the Furin Site. Preprints 2021, 2021020264. doi:10.20944/preprints202102.0264.v1.

5. Antonio R. Romeu, Enric Ollé. The SARS-CoV-2 arginine dimers, 02 August 2021, PREPRINT (Version 1) available at Research Square [<https://doi.org/10.21203/rs.3.rs-770380/v1>].
6. Antonio R Romeu. Probable human origin of the SARS-CoV-2 polybasic furin cleavage motif. BMC Genom Data 24:71, 2023. PMID: 37990144. doi:10.1186/s12863-023-01169-8.
7. Antonio R. Romeu. Synonymous base substitution in SARS-CoV-2 EE.2 lineage furin arginine CGG–CGG codons, 05 July 2023, PREPRINT (Version 1) available at Research Square [<https://doi.org/10.21203/rs.3.rs-3139964/v1>].
8. Antonio R. Romeu. SARS-CoV-2 CQ.1 and CQ.1.1 lineage isolates: 100% synonymous base substitution at furin arginine CGG–CGG codons, 13 July 2023, PREPRINT (Version 1) available at Research Square [<https://doi.org/10.21203/rs.3.rs-3165901/v1>].
9. Antonio R. Romeu. SARS-CoV-2 XBB.1.16.20 lineage study sample: 21% shows a synonymous base substitution in the S gene arginine codons CGG–CGG encoding the S protein furin cleavage site, 18 September 2023, PREPRINT (Version 1) available at Research Square [<https://doi.org/10.21203/rs.3.rs-3356724/v1>].
10. FURIN furin, paired basic amino acid cleaving enzyme [*Homo sapiens* (human)]. NCBI. <https://www.ncbi.nlm.nih.gov/gene/5045> Accessed March 1, 2025.
11. Sun Tian, Qingsheng Huang, Ying Fang, Jianhua Wu. FurinDB: A database of 20-residue furin cleavage site motifs, substrates and their associated drugs. J. Mol. Sci. 8;12(2):1060-5, 2011. PMID: 21541042. doi:10.3390/ijms12021060.
12. Cody B. Jackson, Michael Farzan, Bing Chen, Hyeryun Choe. Mechanisms of SARS-CoV-2 entry into cells. Nat. Rev. Mol. Cell. Biol. 23(1):3-20, 2022. PMID: 34611326. doi:10.1038/s41580-021-00418-x.
13. Sergey A Shiryayev, Andrei V Chernov, Vladislav S Golubkov, Elliot R Thomsen, Eugene Chudin, Mark S Chee, Igor A Kozlov, Alex Y Strongin, Piotr Cieplak. High-resolution analysis and functional mapping of cleavage sites and substrate proteins of furin in the human proteome. PLoS One 8(1):e54290, 2013. PMID: 23335997. doi:10.1371/journal.pone.0054290.
14. Yuelong Shu, John McCauley. GISAID: Global initiative on sharing all influenza data – from vision to reality. Euro. Surveill. 2017;22:30494, 2017. PMID: 28382917. doi:10.2807/1560-7917.ES.2017.22.13.30494.
15. Global Initiative on Sharing Avian Influenza Data (GISAID) database. <https://www.gisaid.org/>. Accessed March 1, 2025.
16. Andrew Rambaut, Edward C Holmes, Áine O'Toole, Verity Hill, John T McCrone, Christopher Ruis, Louis du Plessis, Oliver G Pybus. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5:1403-7. PMID: 32669681. doi:10.1038/s41564-020-0770-5.
17. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J. Comput. Biol. (1-2):203-14, 2000. PMID: 10890397. doi:10.1089/10665270050081478.
18. 18 NCBI BLASTn <https://blast.ncbi.nlm.nih.gov/BlastAlign.cgi>. Accessed March 1, 2025.
19. Fábio Madeira, Young Mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian R N Tivey, Simon C Potter, Robert D Finn, Rodrigo Lopez. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 47(W1):W636-W641, 2019. PMID: 30976793. doi:10.1093/nar/gkz268.
20. EMBL-EBI. Clustal Omega. <https://www.ebi.ac.uk/jdispatcher/msa/clustalo>. Accessed March 1, 2025.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.